PATENT ABSTRACTS OF JAPAN

(11)Publication number : 08-077010
(43)Date of publication of application : 22.03.1996

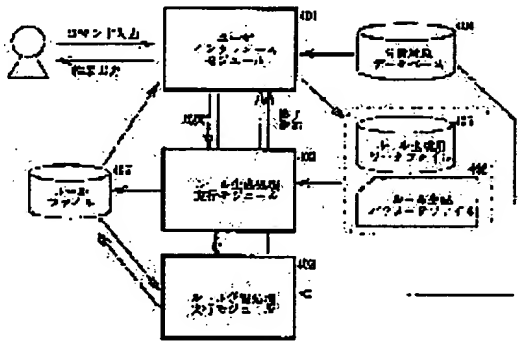
(51)Int.Cl. G06F 9/44
G06F 17/60
G06F 19/00

(21)Application number : 06-239437
(22)Date of filing : 07.09.1994

(71)Applicant : HITACHI LTD
(72)Inventor : MAEDA AKIRA
ASHIDA HITOSHI
TANIGUCHI YOJI
ITO YUKIYASU
TAKAHASHI YORI

(54) METHOD AND DEVICE FOR DATA ANALYSIS

(57)Abstract:
PURPOSE: To provide excellent operation environment wherein a user can generate a result that precisely represent features of data without having any special knowledge when extracting regularity and cause-effect relation between data effective for the user in the form of a rule from information stored in a data base.
CONSTITUTION: The user specifies and corrects a parameter required for processing through a user interface module 401 and utilizes a displayed rule generation result. A rule generating process execution module 402 outputs a rule file 407 based upon input data 405 by using the process parameter which is specified or automatically determined according to data. A rule learning process execution module 403 optimizes the rule by using the contents of a data base 404 to be analyzed and a rule file 407 and stores the result in the rule file 407.



LEGAL STATUS

[Date of request for examination] 23.07.1997
[Date of sending the examiner's decision of rejection] 24.03.2000
[Kind of final disposal of application other than the examiner's decision of rejection or application converted registration]
[Date of final disposal for application]
[Patent number]
[Date of registration]
[Number of appeal against examiner's decision of rejection]
[Date of requesting appeal against examiner's decision of rejection]

[Date of extinction of right]

(19)日本国特許庁(JP)　(12)公開特許公報（A）　(11)特許出願公開番号

特開平８－７７０１０

(43)公開日　平成8年(1996)3月22日

(51)Int. Cl.[6]		識別記号	庁内整理番号	F I		技術表示箇所
G06F	9/44	550 N	7737－5B			
	17/60					
	19/00					
				G06F	15/21	Z
					15/30	Z

審査請求　未請求　請求項の数１３　FD　(全２０頁)

(21)出願番号　特願平6-239437

(22)出願日　平成6年(1994)9月7日

(71)出願人　000005108
株式会社日立製作所
東京都千代田区神田駿河台四丁目6番地

(72)発明者　前田　章
神奈川県川崎市麻生区王禅寺1099番地　株式会社日立製作所システム開発研究所内

(72)発明者　芦田　仁史
神奈川県川崎市麻生区王禅寺1099番地　株式会社日立製作所システム開発研究所内

(72)発明者　谷口　洋司
神奈川県川崎市麻生区王禅寺1099番地　株式会社日立製作所システム開発研究所内

(74)代理人　弁理士　矢島　保夫

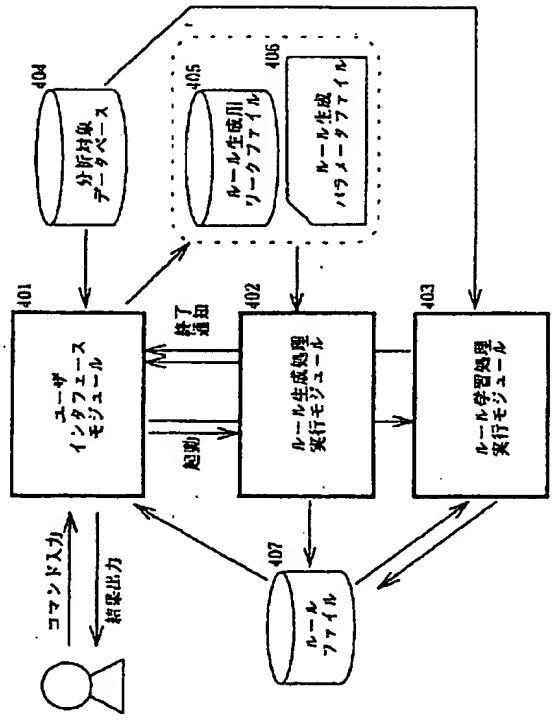
最終頁に続く

(54)【発明の名称】データ分析方法および装置

(57)【要約】　(修正有)

【目的】データベースに蓄積された情報から、利用者にとって有効なデータ間の規則性・因果関係をルールの形式で抽出するのに、利用者が特別な知識を持たなくてもデータの特徴を精度よく表現する結果を生成できる良好な操作環境を提供する。

【構成】利用者は、ユーザインタフェースモジュール４０１を介して処理に必要なパラメータの指定・修正を行い、また表示されたルール生成結果を利用する。ルール生成処理実行モジュール４０２は、指定された、またはデータから自動的に決定した処理パラメータを用いて、入力データ４０５からルールファイル４０７を出力する。ルール学習処理実行モジュール４０３は、分析対象データベース４０４とルールファイル４０７の内容を用いて、ルールの最適化を行い、結果をルールファイル４０７に格納する。



1

【特許請求の範囲】

【請求項１】複数のデータ項目の具体的な値である数値または記号値のデータの集まりを１つのレコードとし、複数のレコードからなるデータを入力して分析することにより、該データの特徴を表現するルールを生成出力するデータ分析方法であって、前記複数のデータ項目中から、ルールの条件節と結論節に用いるデータ項目を選択するステップと、

選択したデータ項目が数値の値を持つデータ項目であるときは、数値を記号値に変換するステップと、

データ項目間の相互関係を項目名称と記号値の組みを１つまたは複数個組み合わせたルール形式で表現した複数の候補ルールを作成するステップと、

該候補ルールが表現する相互関係の強さを評価する評価尺度を算出するステップと、

作成した複数の候補ルールのうち、該評価尺度が大きな候補ルールを１つまたは複数個求めるステップと、

求めた候補ルールを出力するステップとを備えたことを特徴とするデータ分析方法。

【請求項２】複数のデータ項目の具体的な値である数値または記号値のデータの集まりを１つのレコードとし、複数のレコードからなるデータを入力して分析することにより、該データの特徴を表現するルールを生成出力するデータ分析方法であって、前記複数のデータ項目中から、ルールの結論節に用いるデータ項目である結論項目を１つ選択するステップと、

前記複数のデータ項目中から、ルールの条件節に用いるデータ項目である条件項目を選択するステップと、

選択したデータ項目が数値の値を持つデータ項目であるときは、数値を記号値に変換するステップと、

条件項目の項目名称とその記号値の組みを１つまたは複数個組み合わせて条件節を作成し、結論項目の項目名称とその記号値を組み合わせて結論節を作成し、作成した条件節と結論節を組み合わせて複数の候補ルールを作成するステップと、

作成した候補ルールの良さを評価する評価尺度を算出するステップと、

作成した複数の候補ルールのうち、算出した評価尺度が大きな候補ルールを１つまたは複数個求めるステップと、

求めた候補ルールを最終的に生成したルールとして出力するステップとを備えたことを特徴とするデータ分析方法。

【請求項３】前記ルールの条件節に用いるデータ項目を選択するステップは、各データ項目間の関連度を求め、その関連度に基づいて、前記ルールの条件節に用いるデータ項目の選択を自動的に行う請求項１または２のいずれか１つに記載のデータ分析方法。

【請求項４】前記条件項目を選択するステップは、

前記複数のデータ項目中の前記結論項目以外のデータ項

2

目のすべての組み合わせに対して、それぞれ、項目間の関連性の尺度を示す類似度尺度を計算するステップと、

計算した類似度尺度にしたがって、関連性が高い前記結論項目以外のデータ項目を集めて、幾つかのクラスタにクラスタリングするステップと、

前記結論項目と前記結論項目以外のデータ項目とのすべての組み合わせに対して、それぞれ、項目間の関連性の尺度を示す依存度尺度を計算するステップと、

計算した依存度尺度にしたがって、前記クラスタ中で前記結論項目と最も関連性が高いデータ項目を選び、前記条件項目として採用するステップとを備えた請求項２に記載のデータ分析方法。

【請求項５】前記数値を記号値に変換するステップは、幾つかの記号値と前記数値が該記号値に適合する度合を示す記号値ごとのメンバシップ関数とを決定することにより行う請求項１または２のいずれか１つに記載のデータ分析方法。

【請求項６】前記メンバシップ関数値に基づいて、入力されたレコードの重みを算出し、該重みの情報を用いてルールの相互関係の強さを評価する請求項５に記載のデータ分析方法。

【請求項７】分析すべき前記レコード中のあるデータ項目の値が欠損値を含むとき、該データ項目の記号値として欠損値を示す記号値を追加し、生成するルールに該欠損値を示す記号値を使用するか否かを各項目ごとに指定するステップを、さらに備えた請求項１または２のいずれか１つに記載のデータ分析方法。

【請求項８】前記欠損値を示す記号値を使用するか否かの指定を、そのデータ項目と結論項目との依存関係に基づいて自動的に決定するステップを、さらに備えた請求項７に記載のデータ分析方法。

【請求項９】前記候補ルールの評価尺度は、該候補ルールが対象とするレコード数で表現されるカバー率と、該候補ルールの正解率で表現される精度とから算出され、かつカバー率と精度を重要視する度合をパラメータとして指定するステップをさらに備えた請求項１または２のいずれか１つに記載のデータ分析方法。

【請求項１０】入力レコードの前記データ項目間に存在する関数関係を求めるステップをさらに設け、該関数関係に基づいて冗長なルールの生成および表示を行わない請求項１または２のいずれか１つに記載のデータ分析方法。

【請求項１１】出力されたルールの全部または一部と、入力されたデータを用いて、前記ルールの評価尺度を最大化するように、前記メンバシップ関数の形状を指定するパラメータの値を調整するステップを、さらに備えた請求項５または６のいずれか１つに記載のデータ分析方法。

【請求項１２】複数のデータ項目の具体的な値である数値または記号値のデータの集まりを１つのレコードと

し、複数のレコードからなるデータを入力して分析することにより、該データの特徴を表現するルールを生成出力するデータ分析装置であって、前記複数のデータ項目中から、ルールの条件節と結論節に用いるデータ項目を選択する手段と、
選択したデータ項目が数値の値を持つデータ項目であるときは、数値を記号値に変換する手段と、
データ項目間の相互関係を項目名称と記号値の組みを１つまたは複数個組み合わせたルール形式で表現した複数の候補ルールを作成する手段と、
該候補ルールが表現する相互関係の強さを評価する評価尺度を算出する手段と、
作成した複数の候補ルールのうち、該評価尺度が大きな候補ルールを１つまたは複数個求める手段と、
求めた候補ルールを出力する手段とを備えたことを特徴とするデータ分析装置。
【請求項１３】複数のデータ項目の具体的な値である数値または記号値のデータの集まりを１つのレコードとし、複数のレコードからなるデータを入力して分析することにより、該データの特徴を表現するルールを生成出力するデータ分析装置であって、前記複数のデータ項目中から、ルールの結論節に用いるデータ項目である結論項目を１つ選択する手段と、
前記複数のデータ項目中から、ルールの条件節に用いるデータ項目である条件項目を選択する手段と、
選択したデータ項目が数値の値を持つデータ項目であるときは、数値を記号値に変換する手段と、
条件項目の項目名称とその記号値の組みを１つまたは複数個組み合わせて条件節を作成し、結論項目の項目名称とその記号値を組み合わせて結論節を作成し、作成した条件節と結論節を組み合わせて複数の候補ルールを作成する手段と、
作成した候補ルールの良さを評価する評価尺度を算出する手段と、
作成した複数の候補ルールのうち、算出した評価尺度が大きな候補ルールを１つまたは複数個求める手段と、
求めた候補ルールを最終的に生成したルールとして出力する手段とを備えたことを特徴とするデータ分析装置。
【発明の詳細な説明】
【０００１】
【産業上の利用分野】本発明は、データベース装置などの情報記憶装置内に格納された数値または記号で表現されたデータの集まりを分析し、利用者にとって有用な表現に加工変換するデータ分析方法および装置に関する。
【０００２】
【従来の技術】計算機技術の進歩により、計算機内に蓄積されるデータ量は年々増大している。特に、ネットワーク化が進むにつれて、オンラインシステムを中心にこの傾向はますます顕著になっている。現在では、レコード数で１００万件、データ量でギガ（＝１０の９乗）バイトを越えるものも珍しくない。
【０００３】計算機内に蓄積されたデータは、それだけでは数値や記号の集まりに過ぎないことが多い。そこで、このデータの集まりを利用者にとって有用な情報に変換し、データの有効活用を図ろうとする技術が提案されている。最も広く知られている手法は、相関分析・重回帰分析といった統計的手法である。
【０００４】また、比較的新しい手法としては、利用者にとって理解しやすい「もし～ならば…である」といったルール形式に変換して結果を出力する方法、すなわちルールインダクションと呼ばれる知識獲得手法を用いる方法が、知られている。例えば、日立クリエイティブワークステーション２０５０（商品名）ＥＳ／ＴＯＯＬ／Ｗ－ＲＩ解説／操作マニュアルのページ２３～３１には、データ間に存在する関係をルールの形で表現する手法が述べられている。
【０００５】この方法は、本来は与えられたデータからエキスパートシステムへ入力可能なルールを生成することを目的としたものであった。しかし、蓄積されたデータに含まれている因果関係や規則性といった特徴を、人間である利用者が発見するという目的に利用することが可能である。
【０００６】
【発明が解決しようとする課題】上記従来技術は、あくまで計算機が利用できるルールを生成することを目的としている。したがって、人間である利用者が、そのルールを解釈することは可能ではあるが、もとより人間に理解しやすい形式でルールを生成する訳ではないので、人間が解釈しデータの特徴を理解するのに有効なルールを生成することはできなかった。例を用いて、上記文献に述べられている方法を説明する。
【０００７】まず、データを個々の事例の集合と考える。例えば、半導体製造プロセスにおける品質管理データベースを用いて、半導体の不良要因の分析をするといった利用方法では、個々の事例はウェハとよばれる製造単位で管理され、それぞれの製造工程における処理パラメータや、各種検査結果などの情報の組を一つの事例として扱うことができる。図１に、このようなデータの例を示す。
【０００８】また、銀行の顧客データベースから顧客の金融商品の購入動向を調べるといった利用方法では、年齢・預金残高・職業・年収・金融商品購入履歴といった顧客ごとの情報の組が１つの事例であり、分析の対象とするデータはこの事例の集まりと見なすことができる。この例に関しては、実施例で詳しく説明する。
【０００９】上記従来技術によるルール生成の例を説明する。例として、ある金融商品（商品Ａとする）を購入した顧客に共通の特徴を調べることを考える。この場合、従来技術では、データの各項目（年齢や預金残高など）の値から、金融商品Ａを購入した顧客に対応する事

例と購入していない顧客に対応する事例とをできるだけ精度よく分類するルールを生成することが目的となる。

【００１０】上記従来技術では、項目の値の組（例えば年齢が４０歳以上でかつ預金残高が１０００万以上、など）の中で、与えられたデータをもっとも精度よく分類する組を生成する。この場合の精度とは、特定の値を持つ事例の部分集合の中で、金融商品Ａを購入した顧客に対応する事例の割合が大きければ大きいほど、購入した顧客の特徴を精度よく分類するものと考える。この値の組は、「IF　年齢が４０歳以上　AND　預金残高が１０００万以上　THEN　金融商品Ａを購入」というルールの形式で表現することができる。

【００１１】次に、生成されたルールで説明される事例を、全体の事例の集合から除く。上記の例では、年齢が４０歳以上で預金残高が１０００万以上という条件を満たす事例を除くことになる。そして、残された事例の集合に対して、もっとも精度よく分類する項目の値の組を求める。以上の処理を繰り返すことにより、金融商品Ａを購入した顧客を、購入していない顧客から識別するためのルール群を得ることができる。

【００１２】以上の説明から分かるように、上記従来技術で得られるルール群は、

IF 年齢が４０歳以上 AND 預金残高が１０００万以上 THEN 金融商品Ｂを購入

ELSE IF 職業が自営業 AND 年収が８００万以上 THEN 金融商品Ｂを購入

ELSE IF …というように、IF … ELSE IF … ELSE IF …の形式となる。

【００１３】計算機がこのルール群を利用して分類をする場合には、先頭のIFから順に調べていくだけで機械的に処理を実行できる。しかしながら、ルールの数が増えれば増えるほど、人間がこのルールから金融商品Ａを購入した顧客の特徴を理解することは困難になっていく。さらに、１つルールを生成する毎に、残された事例の集合からルールを探しにいくという処理を繰り返すため、事例の数が大きくなるにつれ、必要な処理時間が急速に増大することも問題である。

【００１４】さらに重大な問題として、上記の例題のような実世界のデータの場合、データには一般に非常に大きな雑音が含まれているものと考えなければならない。つまり、金融商品Ａを購入するかどうかというのは、データベースに含まれている項目からは本質的には決定できないはずであるから、もともと非常に精度のよい分類ルールを生成することは期待できない。また、上に挙げた半導体の不良要因分析の場合では、不良の発生がランダムに変動する要因の影響を受けるため、データには大きな雑音が含まれる。このような場合にも、確定的なルールを生成することを望むことが困難な場合が多い。

【００１５】上記のような問題の場合には、データの大体の特徴を表現するような分析方法が有効である。しかしながら、上記従来技術のような方法では、多くの項目の値を組み合わせて少しでも分類精度のよいルールを探しにいくので、一般にIF部にあらわれる条件の数は増大するが、そのルールにあてはまる事例の数は減少するという現象が起こる。これではデータの大体の特徴を理解するという目的を満足することは困難である。

【００１６】実際のデータベースには非常に多くの種類の情報が格納されている。上記の半導体品質管理データの例でいえば、ウェハ番号・製造開始年月日など、また顧客データの例でいえば、氏名・電話番号など、明らかに分析の目的とは関係ないものも含まれている。一方で、半導体品質管理データの場合には製品種別コード、顧客データの場合には住所のように、場合によっては分析有効な情報もある。

【００１７】そこで、そのように多くの種類の情報からなるデータを用いて上記の例のような分析をする場合には、まずどのような項目のデータを使って分類を試みるかということを、利用者があらかじめ指定しなければならない。項目の種類が増えれば増えるほど、この作業は複雑になる。もしかすると関係があるかもしれないような項目を全て含めて分析しようとすれば、必然的に使用する項目数は増え、処理時間の増大を招く。したがって効率よく分析をしようと思えば、注意深く使用項目を選ばなければならず、利用者のノウハウの程度によって、分析の結果の質が大きく左右されることになる。

【００１８】さらに、項目によってはいろいろな見方での分析が必要になる場合がある。例えば、顧客データの例で、地域別に商品購入の動向が異なるような場合、住所という項目を県別で考えるべきか、東北地方／関東地方といったレベルで考えるか、東日本／西日本の２分類で考えるべきかというのは、基本的には分析して見なければ分からない。このような場合を全て尽くそうと思えば、何度も繰り返して従来技術による分析を試みる必要があり、利用者の負担が増大するという問題点がある。

【００１９】これを避けるために、いろいろなレベルでの見方を全てデータ項目に追加するという方法もある。住所の例でいえば、県別／地域別／東西別という項目を分析すべきデータ項目として考えることは可能である。しかし、従来技術による分析では、項目間の意味的な関連は一切考慮していないため、処理に大きな無駄が生じる。

【００２０】例えば、県別というレベルでの分類を試みている間は、それより上位のレベルである地域別や東西別という項目の値は本来考慮する必要がないが、それにもかかわらず、従来技術では無駄な分析を行ってしまう。さらに、項目の処理順序によっては、「IF　住所が関東地方　AND　住所が神奈川県　THEN　…」といった、明らかに意味的に冗長なルールが生成される可能性がある。計算機による分類が目的の場合には、このような冗長性は分類精度には影響しないが、人間がデータの

特徴を理解するという目的には明らかに有害である。

【００２１】また、実際のデータベース中には欠損値と呼ばれる未知のデータ項目が含まれていることが多い。統計手法などで分析をするときには、欠損値は単に無視し、データがないものと見なさざるを得ない。上記ルールインダクションの方法でも、欠損値をもつデータ項目は分類精度には影響しないので、分類ルール中には現れることがない。

【００２２】しかし、データの値が欠損していること自身に意味がある場合がある。例えば住所という項目が欠損値ということが、匿名で銀行口座を作っていることを意味する場合、このことが金融商品Ａの購入に影響を与えることが考えられる。このような場合には、「IF　住所が欠損値　THEN　…」というルールが意味を持つことになる。従来の方法では、このようなルールは一般的には生成できず、また生成するためには欠損値を明示的に特定の値に変換するような人手による処理が必要になるという問題点があった。

【００２３】さらに、上記従来技術による方法では、あくまで分類精度を優先してルールを生成するため、できるだけ多くの事例を説明するという意味で一般的なルールが必ずしも先に生成されるとは限らない。一方、今考えているようなデータ分析方法では、処理時間が長い場合に途中で利用者が割り込みをかけることもある。このようなときに、利用者にとって利用価値の高い一般的かつ単純なルールから先に生成されていれば、割り込みをかけた時点までに生成されたルールを利用することが可能になるが、従来技術による方法ではこのような利用ができないという問題点もある。

【００２４】本発明の目的は、データベースなどに格納されたデータを分析して利用者にとって有用な表現に変換して出力する方法および装置において、上述した問題点を解決することにある。また、本発明の目的は、対象となるデータの量と性質に応じて変換処理方法の一部を自動的に決定することにより、利用者が処理方法に関する特別な知識を有しない場合にも、データの特徴を精度よく表現する結果を生成し、かつ上記の処理を高速に実行し、さらには対話的な計算機環境で上記処理を実行する場合に利用者にとって良好な操作環境を提供するのに適したデータ分析方法および装置を提供することにある。

【００２５】

【課題を解決するための手段】本発明は、複数のデータ項目の具体的な値である数値または記号値のデータの集まりを１つのレコードとし、複数のレコードからなるデータを入力して分析することにより、該データの特徴を表現するルールを生成出力する際に、前記複数のデータ項目中から、ルールの条件節と結論節に用いるデータ項目を選択し、選択したデータ項目が数値の値を持つデータ項目であるときは、数値を記号値に変換し、データ項目間の相互関係を項目名称と記号値の組みを１つまたは複数個組み合わせたルール形式で表現した候補ルールを複数作成し、該候補ルールが表現する相互関係の強さを評価する評価尺度を算出し、作成した複数の候補ルールのうち該評価尺度が大きな候補ルールを１つまたは複数個求め、求めた候補ルールを出力することを特徴とする。

【００２６】また本発明は、複数のデータ項目の具体的な値である数値または記号値のデータの集まりを１つのレコードとし、複数のレコードからなるデータを入力して分析することにより、該データの特徴を表現するルールを生成出力する際に、前記複数のデータ項目中から、ルールの結論節に用いるデータ項目である結論項目を１つ選択し、前記複数のデータ項目中から、ルールの条件節に用いるデータ項目である条件項目を選択し、選択したデータ項目が数値の値を持つデータ項目であるときは、数値を記号値に変換し、条件項目の項目名称とその記号値の組みを１つまたは複数個組み合わせて条件節を作成し、結論項目の項目名称とその記号値を組み合わせて結論節を作成し、作成した条件節と結論節を組み合わせて複数の候補ルールを作成し、作成した候補ルールの良さを評価する評価尺度を算出し、作成した複数の候補ルールのうち、算出した評価尺度が大きな候補ルールを１つまたは複数個求め、求めた候補ルールを最終的に生成したルールとして出力することを特徴とする。

【００２７】以下、本発明を具体的に説明する。まず、分析の対象となる項目（結論項目と呼ぶ）を指定する手段を設ける。上記従来技術で説明した例の場合は、「金融商品Ｂの購入履歴」という項目が結論項目となる。それ以外の項目を条件項目と呼ぶ。

【００２８】次に、対象とするデータ中に含まれる事例の数を調べ、その数に応じて生成されるルールに含まれる条件節数の最大値、各ルールの最小カバー率、ルール群のカバー率、生成するルール数、を決定する。条件節数とは、ルールのIF部に含まれる述語（「アルミ一層配線ショート発生率が５％以上」「年齢が４０歳以上」など）の数である。あるルールのカバー率とは、ルールのIF部を満足する事例の数の、全体の事例の数に対する割合である。ルール全体のカバー率とは、生成されたルール群に含まれるルールのいずれかの条件節を満足する事例の数の、全体の事例の数に対する割合である。

【００２９】次に、データ中の各項目に関してその値が記号属性を持つか数値属性を持つかを調べる。数値属性を持つ項目に関しては、数値の範囲をあらかじめ定められた方法で複数の区間に分割し、各区間に異なる記号値を割り当てることにより、数値属性を記号属性に変換する手段を設ける。分割する区間の数は、あらかじめ定められた値の範囲で、結論項目との依存関係がもっとも大きくなるような数を選択する手段を設ける。この時に欠損値は、あらかじめ定めた特別の値として表現するもの

とする。

【００３０】上記分割する区間には、それぞれメンバシップ関数を割り当てる手段を設け、数値を記号値に変換する時の情報量の損失を小さくする。

【００３１】記号属性を持つ項目に関しては、各項目のとり得る記号値の総数を調べる手段を設け、その数があらかじめ決められたしきい値よりも小さい場合にのみ条件項目とし、それ以外の場合にはルール生成には用いないような処理手段を設ける。

【００３２】さらに、記号属性を持つ２つの項目の対に関し、記号値の間の依存関係が１対多の関係を持つ場合には、１に対応する項目が多に対応する項目の上位概念であるという性質を記憶しておく手段を設ける。この記号属性間の上位概念／下位概念という性質は、ルール生成に使用するデータをＲＤＢなどのデータベースシステムにおいて複数の関係テーブルから定義する場合には、要素テーブル間の関係定義から自動的に抽出する手段を設けることによって得ることもできる。

【００３３】また、数値属性を前述した手段により記号属性に変換した項目に関しても、上記の性質を調べることにより、上位概念／下位概念の性質を求めることも可能である。

【００３４】次に、条件項目として選択された項目のそれぞれに関して欠損値の有無を調べ、欠損値が存在する場合には欠損値であるかどうかが結論項目の値に影響を与えるかどうかを統計的な検定手法などにより調べ、影響があると判断した場合には欠損値をルールの条件部に利用できることを示すあらかじめ定められた別の記号値に変換する手段を設ける。

【００３５】上記の処理の結果をディスプレイなどの表示装置に表示し、表示された情報を利用者が必要に応じて修正できる手段を設けておく。

【００３６】ルール生成としては、カバー率とヒット率から決まるルールの評価尺度を定義し、最大条件節数以下の条件節をもち、かつカバー率が所定の値以上の値をもつ全てのルール候補の中で、大きな評価尺度をもつルールから上記の手段により定められたルール数を選択し、結果として出力装置に出力する手段を設ける。ここで、ヒット率とは、（ルールの条件節と結論節の両方を満たすレコードの数）÷（ルールの条件節を満たすレコードの数）であるものとする。または、評価尺度の大きな順に候補ルールを選択していき、ルール群としてのカバー率が所定の値以上になるまでルールを選択し、生成ルール群とする手段を設ける。

【００３７】前記評価尺度は、ルールのカバー率とヒット率と条件節数から算出する手段を設けておくこともできる。

【００３８】数値属性を記号属性に変換するときにメンバシップ関数を用いた場合には、カバー率およびヒット率は各事例のメンバシップ値を用いて算出する手段を設けておく。

【００３９】上記ルール生成では、条件節数の小さな候補ルールから順に生成し、生成済みの候補ルールに新しい条件節を追加することにより新しい候補ルールを生成する手段を設ける。さらに、生成した候補ルール数に応じてルール生成処理の進行状況を表示装置に表示する手段を設ける。また、利用者が処理の中断を指定する手段を設けるとともに、処理の中断が指定された場合には、その時点までで生成された候補ルールの全部または一部を出力装置に出力する手段を設けておく。

【００４０】ある候補ルールに対して条件節を追加した場合にとりうるルール尺度の最大値を算出する手段を設け、その最大値に応じて条件節を追加するかどうかを判定する手段を設ける。

【００４１】ルール生成手段の中間結果または最終結果として得られたルール群に関し、ルール群全体、または１つまたは複数の選択されたルールに関するルール尺度を最大化するように、数値属性を記号値に変換する際に設けたメンバシップ関数の形状を変更する手段を設ける。

【００４２】

【作用】本発明によれば、対象とする事例の集合から、条件節数の最大値・各ルールの最小カバー率・ルール群のカバー率・ルールの条件部に使用される条件項目・数値属性から記号属性への変換方法・記号属性間の上位概念／下位概念関係・欠損値の扱い方法など、データ分析処理のパラメータを全て自動的に決定することができる。

【００４３】したがって、データを分析しようとする利用者は、説明しようとする結論項目さえ指定すれば、あとは計算機処理によってルール形式で表現されたデータの特徴を結果として得ることができるため、分析作業の手間を軽減し、また利用者のノウハウの存在を前提とすることなく、精度のよいデータ分析方法を提供することが可能になる。もちろん、自動的に決定された上記パラメータを表示装置に表示し、利用者が確認し必要に応じて修正する手段を設けることにより、利用者のデータ分析の目的に、より適した処理を行うことが可能である。

【００４４】また、ルールの良さを評価する尺度として、カバー率・ヒット率・条件節数から決まる評価尺度を用いることにより、利用者がデータの特徴を理解するのにより適したルールが生成される。

【００４５】数値属性を記号属性に変換する際にメンバシップ関数を用いる方法では、例えば同一の区間に属する値でも、その区間内の中心近くの値を持つのか、区間の端に近い値を持つのかがメンバシップ関数の値として区別できるため、記号化することによる情報量の損失を抑えることができ、生成されるルールの精度が向上し、かつ生成されるルールとしては記号値で表現された理解しやすいルールを生成することができる。

【００４６】また、候補となるルールの生成を条件節数の小さなルールから順に行うため、事例の数や項目数が非常に大きくて処理が利用者の想定した時間内に終了せず、利用者の指示によって処理が中断された場合でも、その時点までに生成されたルールがデータの特徴の概要を理解するのに有効なものになる。

【００４７】また、順次候補ルールを生成する際に、条件節を追加して生成されるルールの尺度の最大値をあらかじめ評価しておくことにより、無駄な候補ルールの生成を避けることができ、効率的な処理が実現できる。結果として、処理を高速化することができるようになる。

【００４８】さらに、途中結果または最終結果として得られた（候補）ルール群に含まれるメンバシップ関数で表現された記号値を、ルール尺度が最大になるようにメンバシップ関数の形状を調節することによって、ルールに含まれる記号の意味を、データの特徴をもっともよく説明するものとすることができる。例えば、当初の処理で年齢という項目に対して「３０から３９」という区間を設定したとし、金融商品の購入する顧客が何らかの理由で３３歳から３７歳の区間に集中した場合でも、その特定の区間を自動的にメンバシップ関数で表現できるようになる。つまり、その区間が特徴のある顧客セグメントの存在を示すという情報を利用者に対して提示することができるようになる。

【００４９】

【実施例】以下、図面を用いて本発明の実施例を説明する。本実施例では、前述の金融商品購入顧客の分析の例を用いる。

【００５０】図２は、分析の対象とするデータの例を示す。ある金融商品の購入動向に関して動向の分析をすることを考える。図２のデータには、顧客番号・氏名・年齢・性別・住所・支店コード・県コード・地域コード・預金残高・ローン残高・クレジットカード種別などの情報とともに、ある金融商品の購入実績が格納されている。

【００５１】このデータを分析しようとする利用者は、顧客のどんな属性が金融商品の購入に影響しているかを分析することにより、ダイレクトメールや訪問販売など、その金融商品拡販のための戦略（どのような顧客にアプローチをするのが最適か、など）を検討することを目的としている。

【００５２】図３に、分析処理の全体フローを示す。まず処理３０１では、分析対象データを定義し、図２のような分析対象データの表を作成する。次に処理３０２で、分析対象データの読み込みを行う。

【００５３】処理３０３では、まず「商品購入」という項目が結論項目であることを指定する。次に、データの各項目を調べ、項目が記号属性を持つか数値属性を持つかを調べる。各項目について、条件項目として選択するかどうかを判定し、選択された項目が数値属性をもつ項目である場合には、それを記号値に変換する。最終的に、ルール生成処理の入力となるデータを生成する。

【００５４】処理３０４では、事例数などの情報に基づいて、ルール生成処理のパラメータ（最大条件節数、最小カバー率、など）を決定する。処理３０５では、決定したパラメータを表示装置に表示し、必要ならば利用者がパラメータを修正する。処理３０６ではルール生成処理を実行し、処理３０７で結果を出力装置に出力する。

【００５５】なお、図３ではルールの学習処理は省略したが、処理３０７の後、利用者の指示により随時学習処理を実行することができる。

【００５６】図４に、以上の処理を実行するためのシステムのブロック構成図を示す。システムは大きく３つの処理モジュールからなる。利用者とのインタラクションを制御するユーザインタフェースモジュール４０１、分析対象のデータからルールを生成するルール生成処理モジュール４０２、およびメンバシップ関数の学習によって生成されたルールの微調整を実行するルール学習処理実行モジュール４０３、である。

【００５７】モジュール４０１は、利用者からの指示をコマンドとして認識し、指示された方法に基づいて分析対象データベース４０４からデータを入力し、ルール生成処理実行モジュール４０２の入力となるルール生成用ワークファイル４０５を生成し、同時にルール生成処理を制御するためのパラメータ類の情報をルール生成パラメータファイル４０６に出力する。利用者からのルール生成実行コマンドがモジュール４０１に入力されると、モジュール４０１はモジュール４０２を起動する。

【００５８】モジュール４０２が起動されると、ファイル４０５とファイル４０６の情報を読み込み、ファイル４０５のデータに対して、ファイル４０６で指定されたパラメータでルール生成処理を実行し、結果をルールファイル４０７に出力する。ルールファイル４０７には、生成されたルールの情報と、生成されたルールに含まれる述語を定義するメンバシップ関数の情報が格納されている。

【００５９】モジュール４０２の実行が終了すると、制御はモジュール４０１に戻り、生成されたルールをあらかじめ定められた形式で利用者に表示する。利用者は、表示結果に基づいて、以上の処理をパラメータを変更して再度実行するなどの作業を続行することができる。

【００６０】ここで生成されたルールの学習コマンドがモジュール４０１に入力された場合、モジュール４０１は、モジュール４０３を起動する。モジュール４０３が起動されると、まずデータベース４０４とルールファイル４０７の内容を読み込み、ルールファイル４０７に含まれるそれぞれのルールに対してメンバシップ関数の学習を実行する。処理終了後、学習したメンバシップ関数の情報を再びルールファイル４０７に格納した後、制御をモジュール４０１に戻す。

【００６１】図３に示した処理のうち、処理３０６がモジュール４０２で実行され、残りの処理はモジュール４０１で実行される。

【００６２】以下、各処理の詳細を、図３のフローにしたがって順に説明する。

【００６３】図５および図６は、処理３０１において実行される分析対象データ定義処理の内容を説明する図である。分析の対象となるデータは、複数のリレーションテーブルに格納されており、それぞれ図５(a)の顧客情報テーブル、図５(b)の支店情報テーブル、図５(c)の商品購入履歴テーブルからなる。図６は、図５の３つのテーブルの相互の関連を指定し、分析対象となる図２のテーブル構造定義を模式的に示した図である。これらの定義は、リレーショナルデータベース管理システム（ＲＤＢＭＳ）で提供される機能を用いて行うことができる。

【００６４】図３の処理３０２では、図２で示した分析対象データをデータベースから読み込み、計算機内のメモリに展開する。データ読み込み時には、各項目の値が数値か記号値かということを判定し、かつ数値の場合には最大値／最小値、記号値の場合には異なる記号の数を各項目毎に求めておく。

【００６５】図７に、図３の処理３０３における条件／結論項目の指定画面を示す。処理３０３では、この図７のように画面表示を行い、ユーザに条件項目と結論項目の指定をさせる。

【００６６】まず、データ読み込み直後にはいずれの項目も選択されておらず、図２の分析対象データの各項目の名称が未使用項目名リストボックス７０１に表示されており、条件項目名リストボックス７０２、および結論項目名テキストボックス７０３の内容は空である。

【００６７】ここで、リストボックス７０１の表示の際には、処理３０２で求めた記号値をもつ属性のうち、異なる記号の数があらかじめ定められた数、例えば２０個以上の場合には条件項目／結論項目として選択することを不可とし、その項目名が選択不可であることを表示するものとする。これは、記号値の数が事例数に比べて大きな項目を条件または結論項目として指定すると、生成されるルールが細分化されて利用価値が低下することを防ぐためである。例えば、図７の顧客番号・氏名・住所が相当する。図７では、項目名の先頭に「＊」を付加してその項目が選択不可であることを表現しているが、もちろん網掛け表示／淡色表示などを用いても同様の効果が実現できる。このようにすることにより、ルール生成に使用しても意味のない項目を使用することを防止することができる。

【００６８】リストボックス７０１中の（選択可能な）項目をマウス等の入力装置を用いて複数選び、ボタン７０６をクリックすることにより、選択中の項目が条件項目として選択され、リストボックス７０１からその項目名が削除され、リストボックス７０２にその項目名が追加されて表示される。同様に、リストボックス７０１の項目を１つ選択してボタン７０９をクリックすることにより、結論項目が選択される。リストボックス７０２および７０３の項目を削除するときも同様に、削除項目を選択してボタン７０７、７１０をクリックする。

【００６９】ボタン７０５をクリックすることにより、リストボックス７０１に表示されている項目で選択不可の項目以外を全て条件項目に追加することができる。ボタン７０８は、逆にリストボックス７０２中の全ての項目を条件項目から削除するためのボタンである。

【００７０】これらの選択が終了した後、ボタン７１１をクリックすると、条件／結論項目選択処理が終了する。また、ボタン７１２をクリックすると、その時点での項目選択情報を全て破棄する。

【００７１】次に、ボタン７０４をクリックしたときに実行される条件項目自動選択処理について説明する。ただし、この処理は結論項目が指定されているときに実行できるため、ボタン７０４は結論項目が指定されているときだけ有効であるものとする。

【００７２】図８は、条件項目自動選択処理の処理フローである。図７において、結論項目（テキストボックス７０３）の指定の後、ボタン７０４をクリックすると図８の処理が実行開始する。

【００７３】処理８０１では、使用不可の項目を除く全ての項目に関して処理８０２、８０３を繰り返す。処理８０２では、現在処理している項目が数値属性を持つか記号値属性を持つかを調べ、数値属性を持つ場合には処理８０３を実行する。

【００７４】処理８０３では、処理中の項目の値をあらかじめ定められた個数、例えば５個の記号値に変換する。図２中の「年齢」という項目の例では、項目値のとりうる値の範囲を例えば「２０未満」「２０以上３０未満」「３０以上４０未満」「４０以上５０未満」「５０以上」の５つの範囲に分割し、５つのカテゴリに分類することになる。この分類の方法には、いろいろな方法があるが、例えば、下記の３つの方法などがある。

【００７５】1)等範囲分割：とりうる値の範囲を等分割する。例えば年齢の最小値が０、最大値が７５とすれば、０～１４、１５～２９、３０～４４、４５～５９、６０～７５と分類する。

【００７６】2)平均／標準偏差分割：年齢の値の分布の平均と標準偏差を求め、その値を元に分割する。例えば平均をμ、標準偏差をσとすると、$\mu-0.84\sigma$、$\mu-0.25\sigma$、$\mu+0.25\sigma$、$\mu+0.84\sigma$という値を分割点とする。この値は、分布が正規分布をすると仮定したときに、各分類に含まれる確率を等しくする値である。

【００７７】3)等数分割：実際の年齢の値の分布を調べ、各分類に含まれる事例の数が等しくなるように分類する。

【００７８】どの方法を用いて分類するかは、あらかじ

めデフォルトの方法を用意しておき、必要に応じてユーザが指定できる手段を用意しておくものとする。

【００７９】上記の方法によって分類された数値データは、便宜的に「カテゴリ１」「カテゴリ２」～「カテゴリ５」という値を持つ記号値データに変換される。

【００８０】処理８０４では、結論項目として指定された項目以外の項目中の任意の２つの項目の組からなるすべての組み合わせについて、２つの項目間の類似度尺度をすべて計算する。類似度尺度は、次のように計算する。

【００８１】まず、２つの項目をそれぞれX、Yとし、＊そのとりうる値をxi(i=1,Nx),yj(j=1,Ny)とする。X、Yは、もともと記号値を持つ項目であっても、上記の方法により数値を記号値に変換した項目であってもよい。XとYを確率変数と考えると、その間に相互情報量が定義される。

【００８２】相互情報量に関しては、多くの情報理論の教科書に解説されている。相互情報量とは、例えばX＝xiという情報が与えられたときに、Yの値の発生する確率分布が持つ情報量がどう変化するかに関係した量であり、いわばXとYの間の関連度を示す量である。相互情報量I(X；Y)は、次の式で定義される。

$$I(X;Y)=\Sigma i \Sigma j \; P(xi,yj)\log[P(xi,yj)/P(xi)P(yj)] \qquad (1)$$

【００８３】ここで、Σiは添字iに関する和、P(xi)はX＝xiとなる確率、P(xi,yj)はX＝xi、Y＝jとなる確率を表す。確率はデータ中のレコードの出現頻度から計算するものとし、例えば、P(xi)＝N(xi)／N、Nは全レコード数、N(xi)はX＝xiとなるレコード数とする。

【００８４】I(X；Y)≧０という性質があり、I(X；Y)の値が大きければ大きいほどXとYの間には関連があることを示しているとみなせる。極端な場合、XとYが全く独立の場合には、P(xi,yj)＝P(xi)P(yj)となり、I(X；Y)＝０となる。

【００８５】この性質から、項目Xと項目Yの類似度尺度D(X；Y)を、次の式で定義することができる。

D(X；Y)＝１／I(X；Y)　　(2)

ただし、I(X；Y)＝０のときは、Dの値は非常に大きな固定値をとるものと約束しておく。

【００８６】処理８０５では、処理８０４で結論項目以外のすべての項目の組み合わせごとに計算した類似度尺度Dを基準にして、結論項目以外の項目をクラスタリングする。すなわち、D(X；Y)を項目Xと項目Yの距離とみなし、全ての項目間の距離に基づいて、距離の近い項目をまとめあげ、いくつかのクラスタとして分類する。

【００８７】クラスタリングの方法には様々なアルゴリズムが知られており、その詳細は一般的な統計学の教科書・文献に述べられているので、ここでは説明を省略する。例えば、k-means法と呼ばれる方法を用いるものする。これによって、条件項目の候補となる項目の内、お互いに類似度（関連度）の高い項目を一つのクラスタにまとめあげることができる。

【００８８】多くのアルゴリズムではクラスタ数をあらかじめ指定するが、ここではデフォルト値として全データの項目数の１／１０をクラスタ数とするものとする。すなわち、図２の分析対象データの項目数が６２項目の場合には、６２÷１０＝６（切り捨て）個のクラスタを生成するものとする。もちろん必要に応じてユーザがクラスタ数を直接指定する手段も用意しておくものとする。

【００８９】処理８０６では、結論項目とそれ以外の項目の組み合わせに関して、依存度尺度を計算する。依存度尺度は、例えば処理８０４で説明した相互情報量I(X；Y)を用いることができる。

【００９０】処理８０７では、処理８０５で分類されたクラスタについて、それぞれのクラスタに属する項目の中で最も結論項目との依存度尺度が大きな項目を選択し、それらの項目を条件項目として採用する。

【００９１】以上、図８の処理により、数値・記号値をもつ項目が混在したデータから、結論項目との関連度が最も高く（依存度が大きい）、かつ条件項目間の関連度をできるだけ低く（独立性を大きく）した条件項目を自動的に選択することが可能になる。これによって、以下の処理で生成されるルールの高精度化が図れ、かつ類似の条件項目を使用しないので、ルール間の独立性を高めることができるので、生成ルールの可読性／有用性を高めることができるという効果がある。

【００９２】処理３０３においては、上述の処理で条件項目に選択された項目に対して、欠損値対策処理を実行する。欠損値対策処理は、次のステップを条件項目に選択された各項目に対して行う。

【００９３】ステップ１：選択中の項目に対し、欠損値データが含まれるかどうかを調べる。欠損値データは、特別な数値（例えば負の最大値）または記号値（例えば空の文字列）で表現されているものとする。欠損値が含まれていなければ、処理を終了する。

【００９４】ステップ２：選択中の項目の値が欠損値か欠損値以外かという分類をしたときの、結論項目Zとの相互情報量I(X；Z)を求める。Xは、選択中の項目の値が欠損値か欠損値以外かという値を持つ確率変数である。

【００９５】ステップ３：ステップ２で求めた相互情報量I(X；Z)があらかじめ定められたしきい値Ithよりも大きな場合には、欠損値であるかないかということと結論項目とが所定の程度以上に関連しているということだから、選択中の項目の欠損値を新しい記号値として追加する。

【００９６】以上の処理により、欠損値であること自身に意味があるとステップ３で判断された場合には、「欠

損値」という特別の記号値がその項目に追加され、後述のルール生成にその記号値を使用することができるようになる。それ以外の場合には、その項目の欠損値は、後述するルール生成処理では一切無視される（どの記号値にも属さず、生成ルールにも決して現れることがない）特別な値として扱われる。

【００９７】図３の処理３０３における数値→記号値変換処理は以下のように行う。

【００９８】図９は、数値→記号値変換方法の一覧表示である。ユーザは、表示装置場の図９のような表示を参照して、マウスなどを用いて各項目について変換方法を指定する。

【００９９】図９において、項目名表示欄９０１には、本来数値属性をもち、記号値に変換される項目の名称が表示される。記号数表示欄９０２には、記号値に変換した後の異なる記号の数が表示される。記号名称表示欄９０３には、記号値の名称が表示される。

【０１００】図９の例では、年齢は２０未満、２０代、…、預金残高は少ない、ふつう、多い、ローン残高は小さい、中くらい、…、とそれぞれ異なる記号値が割り当てられている。変換方法表示欄９０４には、図８の処理８０３で説明した数値→記号値変換方法が表示される。図９の例で、年齢に対応する変換方法がユーザ指定となっているが、このユーザ指定とは、ユーザが分割方法を直接指定したことを示している。

【０１０１】各項目に対する記号値変換方法の指定が終了した後、ボタン９０６をクリックすることで変換方法指定処理が終了し、ボタン９０７をクリックすることでそれまでの指定内容が破棄される。ボタン９０５をクリックすると、図８の処理８０３でも説明したようにデフォルトの記号値変換方法で全ての項目の変換方法をリセットする。

【０１０２】各項目毎に変換方法を指定するには、欄９０１の項目名の部分をダブルクリックする。これにより、図１０の変換方法指定画面が表示され、当該項目の変換方法を指定することができるようになる。

【０１０３】図１０において、オプションボタン１００１は、選択中の項目の記号値変換方法として、等数分割・等範囲分割・平均／標準偏差分割・ユーザ指定分割のどの方法を用いるか、およびそれぞれの分割方法をクリスプ的に行うかファジィ的に行うかを指定する。

【０１０４】テキストボックス１００２には、記号値に変換する際の記号数を入力する。

【０１０５】記号名称表示欄１００３には、現在の記号名称が表示されている。ユーザが、この記号名称を変更したい場合には、直接、欄１００３の該当領域をクリックすることにより記号名称を入力することができる。これは、各記号値の最小値表示欄１００４および最大値表示欄１００５に関しても同様である。ただし、最小値または最大値を直接ユーザが変更した場合には、オプションボタン１００１は自動的にユーザ指定分割方法を指定した状態に変化するものとする。

【０１０６】欄１００４または欄１００５の値が変更されると、レコード数表示欄１００６の内容もそれに応じて自動的に変更される。これは、メモリ中に記憶されているデータを用い、それぞれの記号値に対応するレコードが何件あるかをカウントすることによって行う。

【０１０７】オプションボタン１００１のうちクリスプ分割またはファジィ分割のいずれかを指定することによって、欄１００４および欄１００５で表示される最小値／最大値の範囲で記号値変換処理をクリスプ的に行うかファジィ的に行うかを指定できる。ファジィ的に行うことにより、記号値変換によって失われる情報量を最小限にし、生成されるルールの精度を向上することができる。以下、この処理について説明する。

【０１０８】ファジィ的な分割をする際にも、各記号値の値の範囲は図１０の欄１００４および欄１００５には最小値／最大値として表示されるが、それをファジィとして解釈する。図１１は、このようなファジィ的分割を図１０のローン残高の記号値変換の例で説明する図である。

【０１０９】図１１の横軸はローン残高の数値を万円単位で表示し、縦軸は「小さい」「中くらい」「大きい」という（ファジィ）記号値の適合度を表す。「小さい」という記号値にはメンバシップ関数１１０１が、「中くらい」にはメンバシップ関数１１０２、「大きい」にはメンバシップ関数１１０３が対応する。メンバシップ関数の形状は、傾きの絶対値が一定、各記号値の分割点（図では、「小さい」「中くらい」の分割点である５００、「中くらい」「大きい」の分割点である１０００の２カ所）でそれぞれのメンバシップ関数値が0.5になり、かつどのような値に対しても０以外の値をもつメンバシップ関数は高々２個であり、かつメンバシップ関数値の和が1.0になるように設定することができる。

【０１１０】例えば図１１で、「ローン残高＝６００（万円）」という数値は、「中くらい」という記号値に対してが0.8、「大きい」という記号値に対して0.2というメンバシップ値を持つことになる。すなわち、「ローン残高＝６００」という値をもつレコードは、「ローン残高」が「中くらい」というレコードが0.8個分、「大きい」というレコードが0.2個分の２つに分割されたように振る舞うものとみなすことができる。

【０１１１】このようなファジィ分割の場合には、図１０の欄１００６に表示されるレコード数も、それぞれのメンバシップ値の合計と考えることができ、一般に実数の値をとることになる。

【０１１２】次に、図３の処理３０４におけるルール生成パラメータ算出処理を説明する。ルール生成処理を制御するパラメータとしては、次の５種類のパラメータを考える。

・生成ルール数
・全体カバー率
・最大条件節数
・最小カバー率
・カバー率優先係数

【０１１３】生成ルール数は、ルール生成処理で生成すべきルール数を指定するパラメータである。デフォルト値は、使用する条件項目数Ｎから次の関係式にしたがって定める。

【０１１４】（生成ルール数のデフォルト値）＝Ｎ　＊（最大条件節数）

これは、一般的に条件項目数が多くなればデータの特徴を説明するために生成すべきルール数も増えることを反映させたものである。

【０１１５】カバー率とは、あるルールの条件部を満たす事例の数の、全体の事例数に対する割合と定義する。全体カバー率は、生成したルールに対して、少なくとも１つのルールの条件部を満たす事例の数の、全体の事例数に対する割合を指定するパラメータである。生成するルール数を大きくすれば全体カバー率は大きくなるという関係がある。全体カバー率に対してはデフォルト値は用意しない（すなわちユーザの明示的な指定によってのみ値を設定できる）。

【０１１６】最大条件節数は、ルールの条件部に含まれる述語数の最大値を指定するパラメータである。例えば、最大条件節数３に指定した場合、

もし、X1がA1　かつ　X2がA2　かつ　X3がA3　ならばZはB

というルールは生成されうるが、さらに「～　かつ　X4がA4　ならば　～」というルールは（条件節数が４となるので）生成されないことになる。デフォルト値は３とする。

【０１１７】最小カバー率は、生成されるルールのカバー率の最小値を指定するパラメータで、ここで指定された値以下のカバー率をもつルールは、特殊なルールとして生成されない。最小カバー率のデフォルト値は、１００／（生成ルール数）（単位パーセント）とする。これは、一般的に生成ルール数が多いほど細かなルールでも必要とされることを考慮したものである。

【０１１８】カバー率優先係数は、ルールに対する評価尺度におけるカバー率と精度の関係を指定するもので、詳細は後述のルール生成処理において説明する。デフォルト値は1.0（カバー率を最大限に優先）とする。

【０１１９】図３の処理３０５では、処理３０４で算出＊したルール生成パラメータを表示装置に表示し、必要に応じてユーザが修正できるようにする。図１２に、表示画面例を示す。

【０１２０】テキストボックス１２０１～１２０５には、図３の処理３０４で算出されたルール生成パラメータの値が表示される。図の例では、生成ルール数＝２４、全体カバー率＝指定なし、最大条件節数＝３、最小カバー率＝４％、カバー率優先係数＝1.0である。

【０１２１】ユーザは、テキストボックスを直接クリックすることで表示されたパラメータの値を入力することができる。図１２の表示画面には、分析対象となるデータの他の情報（データファイル名、レコード数、条件項目数、結論項目とその記号値）が表示領域１２０６に表示されているので、ユーザはこれらの情報を参考にしながらルール生成パラメータの値を修正することが可能である。

【０１２２】パラメータの修正後、ボタン１２０７をクリックすることにより修正した値が有効になり、ボタン１２０８がクリックされると修正した値は破棄される。

【０１２３】図３の処理３０６では、処理３０４および処理３０５で指定されたルール生成パラメータにしたがってルール生成処理を実行する。

【０１２４】ルール生成処理では次の形式のルールを生成する。

・もし　X1　が　A1　かつ　X2　が　A2　かつ　…　かつ　Xn　が　Anならば　Y　が　B

ここで、Xiはそれぞれ条件項目名称、Aiは項目Xiの記号値の名称である。同様に、Yは結論項目名称、Bは項目Yの記号値の名称である。

【０１２５】「XiがAi」という組を条件節と呼ぶ。上記のルールは、条件節をｎ個もつルールの例である。条件節全体を条件部とよぶ。また「YがB」を結論節と呼ぶ。本実施例では結論部は１つの結論節からなるものとしておく。

【０１２６】ルールには、評価尺度と呼ばれる実数値を割り当てる。評価尺度の大きなルールほど、ルールとしての価値が高いものとみなす。

【０１２７】図３の処理３０６では、条件節と結論節の組み合わせの中から、ルール評価尺度の大きなルールを選び出すという処理を実行する。これにより、価値が高いルールを選ぶことができる。

【０１２８】本実施例において、「もし　A　ならばB」というルールの評価尺度μ（A→B）は次式のように定義される。

$$\mu(A \rightarrow B) = P(A)^{\beta} * \log[P(B \mid A) / P(B)] \quad (3)$$

【０１２９】ここで　a^b　はaのb乗を意味する。Ｐ（Ａ）は分析対象とするデータの中で条件部Ａが満足される確率、すなわち事例全体の中でＡという条件を満たす事例の割合、を表す。同様に、Ｐ（Ｂ）は結論部Ｂが満足される確率、Ｐ（Ｂ｜Ａ）はＡという条件を満たすという前提で結論部Ｂが満足される確率を表す。βは図３の処理３０４で説明したカバー率優先係数である。式（３）を事例の数で書き直すと、Ｐ（Ｂ｜Ａ）＝Ｐ（Ａ＆Ｂ）／Ｐ（Ａ）であることから、

$$\mu\ (A \rightarrow B) = [N(A)/N]^{\beta} * \log[N \cdot N(A\&B)/N(A)N(B)] \quad (4)$$

となる。

【０１３０】ただし、

N:　　全体の事例数

N(A):　条件部Ａを満たす事例の数

N(B):　結論部Ｂを満たす事例の数

N(A&B):　条件部Ａと結論部Ｂを同時に満たす事例の数である。

【０１３１】式(4) の定義の第一因子はルールのカバー率のβ乗であり、第二因子は精度に対応する。すなわち、カバー率が大きく（できるだけ広い範囲の事例を説明し）、かつ精度がよい（ルール全体が成立する確率が大きい）ほど、高い評価尺度をもつことになる。一般には広い範囲をカバーするほど精度は低下する傾向にあるので、これら２つの因子は互いに背反する傾向にある。ルール生成パラメータの一つであるカバー率優先係数βは、評価尺度のなかでどの程度カバー率を重視するかを決めるもので、β＝1.0 の時にはカバー率を最も優先して考慮し、β＝0.0 の時にはカバー率は考慮せず精度だけを考慮してルールを評価することになる。

【０１３２】もちろんここで定義したルール評価尺度は、あくまでも１つの候補であり、いくつかの評価尺度の定義を追加すること、またそれらの定義をユーザオプションとして切り替えて使用すること、などが可能である。

【０１３３】ルール生成処理では、式(4)で定義したルール評価尺度の大きなルールから指定された個数のルールを生成する。探索すべき空間は、すべての可能な条件節の組み合わせになる。

【０１３４】条件項目１０個、それぞれの記号数３程度の問題でも、この探索空間は膨大な広さになる（およそ３０の１０乗のオーダ）ため、全数探索を現実的な時間で実行することは不可能である。したがって、以下に説明する方法を用いて、探索範囲の制限と枝刈りにより探索を効率化する。

【０１３５】まず、最大条件節数と最小カバー率を用いてルール探索範囲が制限できる。

【０１３６】これらの制限後の空間での探索方法として、本実施例では深さ優先探索法を用いる。

【０１３７】図１３に、探索空間のイメージを示す。基本的には条件節述語の組み合わせであるから、木構造の探索手法が使える。

【０１３８】図１３において、例えば、ノード１３０１に対応するルールは、if　X1　isA　then　…であり、ノード１３０２に対応するルールは、if　X1　is　A　andX2　is　B　then　…である。すなわち、子ノードに対応するルールの条件節は、親ノードに対応するルールの条件節に１つ述語を追加したものになる。各ノードの近傍に付された数値は、当該ノードに対応するルールの条件節の述語を示すものとする。例えば、ノード１３０１の近傍には{1} と記載されているが、これは対応するルール中の述語「X1　is　A」を示している。また、ノード１３０２の近傍には{12}と記載されているが、これは対応するルール中の述語「X1　isA　and　X2　is　B」を示している。

【０１３９】図１３では、最大条件節述語数３の場合を示した。木の高さ（ルートから終端ノードまでの最大リンク数）は最大条件節述語数になる。

【０１４０】条件節に現れる述語の順序には意味がないので、例えばノード１３０３の子ノードは{21}でなく、{23}から始まっていることに注意しておく。また１つの条件節に、同じ条件項目で異なる記号値をもつ２つの述語は含まれない。親ノードの条件節に１つ述語を追加して子ノードを生成する時には、この条件を考慮しておく必要がある。

【０１４１】深さ優先探索では、図１３に示すすべてのノードを次の順序で生成し、評価する。

1→12→123→124→…→12n→13→134→135→……→(n-2)(n-1)n→(n-1)→(n-1)n→n

【０１４２】深さ優先探索の場合は、常にそれまでで評価尺度の最も大きなルールＮ個（Ｎは生成ルール数）を保持しながら上記の順序で候補ルールを調べていけばよい。

【０１４３】次に、上述したルールの探索を高速化するための枝刈り方法について説明する。図１３のノード数は前述の例で３０の３乗程度であるから、直接全数探索することも不可能ではないが、数10～数100万件のデータを対象にすること、項目数も１００程度になる場合も少なくないことから、できるだけ効率のよい探索方法を採用する必要がある。

【０１４４】効率向上の基本的な考え方は「枝刈り」である。すなわち、ある中間ノードでの状態を評価することによって、その子ノードの評価をせずにすませることである。本実施例におけるルール生成処理の場合、カバー率による枝刈りとルール評価尺度による枝刈りを行う。

【０１４５】カバー率による枝刈り方法の基本的な考え方は、あるノードに対応するルールを生成した時、そのルールのカバー率がユーザ指定のカバー率下限を下回った場合には、そのノードおよびすべての子ノードを探索する必要がない、ということである。子ノードに対応するルールの条件節は親ノードに対応するルールの条件節に１つ述語を追加したものになるから、親ノードのルールのカバー率が下限を下回ったときは、子ノードのルールのカバー率も必ずその下限を下回ることになる。

【０１４６】ルール評価尺度による枝刈りについて説明する。探索においてＮ個のルールを生成する場合、探索のある時点でのＮ番目の候補ルールの評価尺度をμ_{th}とする。枝刈りの基本的な考え方は、あるノードをルート

とする部分木におけるルール評価尺度の最大値を求め、その値がμthよりも小さな場合には部分木のノードは生成しない、ということである。ここで、いかに部分木に含まれるノードの評価尺度の最大値を推定するかが、枝刈りの効率を決める重要な因子になる。

【０１４７】以下、この評価方法について説明する。簡＊単のため、カバー率優先係数β=1.0の場合で説明する。一般の値の場合での拡張は容易にできるので、ここでは省略する。

【０１４８】あるノードにおけるルールを「もしＡならばＢ」とすると、その評価尺度は次式で定義される。

$$\mu\ (A \rightarrow B) = [N(A)/N] * \log[N \cdot N(A\&B)/N(A)N(B)] \quad (5)$$

【０１４９】条件節Ａにいくつかの述語を追加した条件節をＡ'とすると、μ（Ａ'→Ｂ）の最大値は N(A')=N(A&Bi) のときに限られる。なぜなら、上式でN、N(B)は定数とみなしてよく、また上式はN(A&B)に関して単調増加であるから、N(A&B)も定数とみなしてよい（すなわち※ち、精度をあえて下げるようなルールは決して評価尺度を最大にしない）。したがって、式(5)は下記のように書くことができ、この式について、N(A&B)≦N(A')≦N(A)の範囲で、

※

$$\mu\ (A' \rightarrow B) = [N(A')/N] * \log[N \cdot N(A'\&B)/N(A')N(B)] \quad (5)$$

の最大値を求めればよい。これは、上式をN(A')で微分すると必ず＜０になることから、N(A')=N(A&Bi)のときに最大値 μmax＝[N(A&B)/N]*log[N/N(B)] となる。この値がμthよりも小さければ、そのノードの子孫ノードを探索する必要がない。

【０１５０】実際の探索においては、早い段階で評価尺度の高いルールを見つけておくことが、効率向上の点で有効である。早く見つければ、それだけμthの値が大きくなり、枝刈りが実行される可能性を増やすことができる。

【０１５１】そのための探索戦略の一方法として、図１３の木構造をまともに深さ優先で探索するのではなく、まず条件節述語数１の場合をすべて調べ、条件節述語数≧２の場合を深さ優先探索にすることが可能である。すなわち、β＞０ならばカバー率の大きなルールほど大きな評価尺度をもつから、条件節述語数１の場合を先に調べることで、μthの値を早く大きくすることができる。

【０１５２】このような探索を行うことによって、ルール生成処理を高速化できるだけでなく、ルール生成処理の途中でユーザが処理の中断を指示したときに、中間結果として得られるルールがデータの特徴を全体として表現したルールになる、という効果もある。

【０１５３】また、生成されたルールはユーザにとってできるだけ理解容易なものであることが望ましい。言い換えると、ルール評価尺度に差がなければ、条件節の述語数は少ないほうがよい。これをルール生成に反映させる方法としては、「条件節述語を追加したルールの精度は、追加する前のルールの精度よりも大きくなければならない」という制約を課すことにする。すなわち、条件節述語を追加してルールを複雑にしたのだから、その分ルールの精度は上がらなくてはならない、ということである。この制約をつけ加えてルール生成を行うことにより、よりわかりやすく利用価値の高いルールを生成することができる。

【０１５４】さらに、本実施例のルール生成処理では、もともと記号値属性をもつ条件項目間の１対多の依存関係を考慮してルールを生成する。例えば、図２の県コードと支店コードを考える。支店コードは各支店に固有のコードであるから、支店コードを決めれば県コードが唯一に決まる。一つの県に複数の営業店が存在する場合には、逆は成立しない。このとき、県コードと支店コードの間には１対多の関係が存在するという。

【０１５５】この２つの条件項目を含むデータに対してルール生成を行うと、支店コードと県コードを同時に条件節述語に含むルールが生成される可能性がある。少なくとも、木構造の探索では候補ルールとして評価尺度が計算される。しかし、これは全くの無駄である。支店コードが条件節述語に含まれていれば、そのコードから県コードは唯一に決まるのだから、県コードに関する条件節述語を追加する必要はない。

【０１５６】カテゴリカルデータ間の１対多の関係は、分類概念の上下関係と見なすこともできる。すなわち、県コードが上位の分類概念、支店コードが下位の分類概念になる。

【０１５７】この上下関係は木構造では表現できない。支店コードの例で、さらに地方コードと支店規模コード（大規模・中規模・小規模など）を考える。地方コードは県コードのさらに上位の概念であるから、営業店コードと地方コードの間にも上下関係が存在する。さらに、営業店コードを指定すれば支店規模コードも決まるから、ここにも上下関係が存在する。すなわち、県コードと支店規模コードはどちらも支店コードの上位概念である。逆に、１つのカテゴリカルデータが複数の下位概念を含むこともある（例えば、県コードに対する支店コードと取引先コード、など）。

【０１５８】本実施例では各条件項目に対して、その上位概念に相当する項目のリストを作成しておく。このリストは、例えば図６に示した分析対象データの定義から自動的に作成することができる。すなわち、リレーショナルデータベースにおいては各テーブルの主キーはテーブル内のレコードを一意的に指定できるものでなければならないから、テーブル内の主キーの項目とそれ以外の項目の間には１対多の関係が常に存在する。図６の例では、支店コードと県コード、支店コードと地域コードの

間に１対多の関係がある。もちろんこれらの関係は、図３の処理３０３で説明した条件項目の選択時に、条件項目間の相互情報量を算出したときに、容易に自動的に発見することもできる。

【０１５９】これらの上位概念リストを用いてルールの生成時の木構造の探索を行うときには、追加すべき述語に対応する項目の上位概念または下位概念の項目がすでに条件節に含まれているかどうかをチェックする（直接の上位／下位概念だけでなく、上位の上位／下位の下位にあたる項目などもチェックする）。もし含まれていれば、その項目の述語は追加の対象とはしない。

【０１６０】次に、数値→記号値変換でファジィ分割を用いる場合の処理内容を説明する。

【０１６１】本実施例のルール生成処理では、図１１のファジィ分割方法がもつ特徴を利用する。図１１のファジィ分割では、１つの数値は高々２つの記号値に分配され、それらの和は必ず 1.0 になる（例えば、「小さい」0.3、「中くらい」0.7、「大きい」0.0のように）。また、図４のルール生成用ワークファイル４０５中の表現では、「小」を０に、「中」を１に、「大」を２に割りあてる。このとき、数値 x は次の規則により１つの実数値 y に変換される。

【０１６２】ファジィ記号値変換規則：もし「小さい」の割合が z、「中くらい」の割合が (1-z) ならば、

y = 1-z

もし「中くらい」の割合が z、「大きい」の割合が (1-z) ならば、

y = 2-z

ただし、0≦z≦1 とする。

【０１６３】すなわち、０と１の間の実数値 y は、「小」と「中」という記号の配分比を表すものと約束する。図１１のファジィ分割の特徴から、１つの数値が３つ以上の記号値に分配されることはなく、また和が1.0になることから、この定義で矛盾は生じない。

【０１６４】上記の方法でファジィ的に記号化された事例を (y1, y2, …, yn) とする。ルール生成処理において事例の数を計算するのは、ルールの評価尺度を計算するときである。クリスプな分割のときは、ある条件節を満たす事例がいくつあるかをカウントする処理になる。ファジィ記号化では次のようになる。

【０１６５】x1 が「小さい」に該当する事例は、0≦y1≦1 の値をもつ事例で、その重みは 1-y1 になる。

【０１６６】x1 が「中くらい」に該当する事例は、0≦y1≦2 の値をもつ事例で、その重みは 1-|1-y1| になる。

【０１６７】x1 が「大きい」に該当する事例は、1≦y1≦2 の値をもつ事例で、その重みは y1-1 になる。

【０１６８】もちろん、条件節述語が複数ある場合には、上記の重みをそれぞれの述語に関して掛け合わせたものが全体の事例の重みになる。ルール評価尺度の定義に現れる N(A) や N(A&Bi) などは、これらの事例重みの和として考え、実数値として取り扱うものとする。

【０１６９】上記の処理でファジィ分割により記号値に変換した項目をルール生成処理で使用すれば、分割における境界付近のデータを正しく扱うことができるので、生成されたルールを高精度化することができる。

【０１７０】図４のワークファイル４０５中のデータでは、値－１は欠損値を表すものとする。ルール生成処理には数値属性そのものは使用せず、また記号値は０から順番に割り当てて行くので、この値が混乱を起こすことはない。

【０１７１】欠損値自身をルール生成に使用するかどうかは、図３の処理３０３における欠損値対応処理か、またはユーザが指定した内容で決まるものとする。ルール生成に使用する場合には、「もし x が欠損値で、…」というルールが許される。欠損値をルール生成に使用しない場合には、単に欠損値は無視される（すなわちどの記号にも一致しない記号として扱われる）。欠損値は特別な値なので、前項のファジィ記号化の規則は適用されない。

【０１７２】本実施例では、ルール生成処理がどこまで進んだかを示す進行状況を表示する。ルール生成処理では、本来調べるべきルールの個数はあらかじめ算出できるから、その内の何パーセントを調べたかという数字を、適当なタイミングで表示するものとする。枝刈りなどで処理をスキップされたルールも当然カウントする。

【０１７３】ルール生成処理では、ユーザ割り込みを受け付けるものとする。上述のルール生成進行状況表示に、中断コマンドを受け付けるボタン１４０１を用意しておく。図１４に、ルール生成処理進行状況および中断ボタン１４０１の表示例を示す。

【０１７４】ボタン１４０１がクリックされ処理の中断が指定された場合、次の動作は「続行」か「終了」かのどちらかになる。続行の場合にはそのまま処理を続け、終了の場合にはルール生成処理を終了し、現時点までに生成されたルールを表示するものとする。

【０１７５】図１５に、以上説明したルール生成処理の処理フロー図を示す。候補ルール集合のクリアおよびμthのゼロクリアなどの初期化（１５０１）の後、すべての候補についてステップ１５０３～１５１２を繰り返す。まず、評価尺度μを計算し、μとμthとを比較する（１５０４）。μ＞μthなら、候補ルール集合に当該候補ルールを追加し（１５０５）、μthに候補ルール集合中で最小の評価尺度をセットする（１５０６）。μ≦μthなら、サブノードに関する評価尺度の最大値μmaxを計算し（１５０７）、μmaxとμthとを比較する。μmax＜μthなら、サブノードをすべて評価済みとする（１５０９）。

【０１７６】次に、ルール生成処理進行状況を表示し

（１５１０）、処理中断が指示されたかどうか判定する（１５１１）。中断が指示されていたら、ループから脱出する（１５１２）。すべての候補についてループ処理が終了するか、ステップ１５１２でループから脱出したときは、候補ルール集合を出力して（１５１３）、処理を終了する。

【０１７７】次に、図４のルール学習処理実行モジュール４０３で実行される処理について説明する。

【０１７８】図１６は、学習すべきメンバシップ関数の形状パラメータを示す。本実施例では、学習によって調節されるパラメータは分割点α1とα2とする。すなわち、ルール生成処理で生成されたルールに対して、最もルール評価尺度が大きくなる分割点α1とα2を求める。ただし、ファジィ分割を用いたルール生成処理において、１つの数値は高々２つの記号値に分配され、それらの和は必ず 1.0 になるという性質を仮定しているため、その仮定を満足する範囲でのみα1とα2の値を変化させるものとする。図１６には、学習後のメンバシップ関数形状の例を示した。

【０１７９】具体的な学習は次のようなステップで行う。

ステップ１：学習すべきルールを指定する。どのルールを学習するかをユーザが指定する。

ステップ２：ステップ３～ステップ８を学習終了条件が成立するまで繰り返す。

ステップ３：各パラメータの変更量Δαi＝０（i=1,2,…,m）とする。mは調節すべき分割点パラメータの総数である。

ステップ４：指定されたルールのそれぞれに対してステップ５～ステップ７を繰り返す。

【０１８０】ステップ５：処理中のルールの評価尺度μを算出すると同時に、∂N(A)／∂αi、∂N(B)／∂αi、∂N(A&B)／∂αiを算出する。N(A)は条件部Ａを満足する（ファジィ分割を考慮した）事例の数であり、αiを微小量Δαi変化させたときのN(A)の変化量ΔN(A)は、図１６のメンバシップ関数形状から容易に求めることができる。∂N(B)／∂αi、∂N(A&B)／∂αiの値も同様に求める。

ステップ６：μの定義式(4)とステップ５で求めた値から、∂μ／∂αi＝∂N(A)／∂αi・∂μ／∂N(A)＋∂N(B)／∂αi・∂μ／∂N(B)＋∂N(A&B)／∂αi・∂μ／∂N(A&B)なる関係式を用いて、∂μ／∂αi（i=1,…,m）を求める。

【０１８１】ステップ７：全てのiに対して、Δαi←Δαi＋∂μ／∂αiとする。

ステップ８：全てのiに対して、αi←αi＋λ・Δαiとする。λはあらかじめ定められた学習係数である。

ステップ９：学習済みのαiの値を図４のルールファイル４０７に出力する。

【０１８２】ステップ２における終了条件としては、例えば

・指定された回数ステップ３～ステップ８を繰り返した
・全てのｉに対して｜Δαi｜＜εが成立した（εはあらかじめ定められた正の定数）
・ルール評価尺度μの変化｜Δμ｜＜εが成立した

などの条件、または上記条件の組み合わせ、などを用いることができる。

【０１８３】以上の処理により、１つまたは複数のルールの評価尺度を最大化するようなメンバシップ関数パラメータを学習により自動的に求めることができる。これにより、データ間に存在する規則性・因果関係をより精度よく抽出することが可能になるとともに、規則性を説明するために最適な項目値の分割点自身を最適に決定することができるようになるという効果がある。また、この学習によれば、生成ルール全体またはその部分集合の評価尺度の和を最適化することが可能なため、ユーザにとって意味のあるルールだけを学習によって最適化することができ、より利用価値の高いルールを生成することができるという効果もある。

【０１８４】

【発明の効果】以上説明したように、本発明によれば、次のような効果を実現することができる。

【０１８５】まず、データ間に存在する規則性・因果関係をユーザにとって理解しやすいルールという形で抽出できるため、従来の方法では困難であった大量データの有効活用が可能になるという効果がある。特に、製造プロセスにおける検査データの分析による不良発生要因の解析や、顧客データベースの分析による新商品のマーケティング戦略の立案などにおいて、有効な情報をデータベースから抽出することができるようになるという効果がある。

【０１８６】さらに、データベース中に含まれる項目のうち、ルール生成に使用する項目を自動的に選択でき、またその他の処理パラメータもデフォルト値が用意されているため、データの性質に関して詳細な知識を持たないユーザでも、分析に当たって複雑な前処理をすることなしに、データ分析処理をほぼ全自動的に実行することができ、きわめて使い勝手のよい分析手法を提供することが可能になるという効果がある。

【０１８７】また、データベースに含まれる数値データは、記号値に変換されてルールに使用されるため、雑音の大きなデータに対してもマクロな構造に対応した規則性を発見でき、それだけユーザにとって利用価値の高い結果を生成できるという効果がある。

【０１８８】上記記号値への変換をファジィ的に行うようにすれば、記号値へ変換することによる情報の損失を最小限に押さえ、精度のよいルールを抽出できるという効果がある。

【０１８９】また、データ中に欠損値が含まれる場合でも、その欠損値に意味があるとしてルールに使用するこ

とも、また、単に未知のデータとして扱うこともできるため、欠損値の含まれるデータでも有効に活用できるという効果がある。また、これらの判断を自動的に実行することもできるため、より容易に分析処理が実行できるといういう効果もある。

【０１９０】ルールの生成処理においては、より複雑なルールを生成した場合のルール評価尺度の最大値を評価しながら実行するため、無駄な候補ルールの生成をさけることができるため、より高速にルール生成ができるという効果もある。

【０１９１】ルール生成処理パラメータは、データ数やデータ項目数に応じて最適な値を自動的に生成するため、ユーザは細かなパラメータの指定に煩わされることなく処理が実行できるという効果もある。

【０１９２】ルール評価尺度では、カバー率と精度の両方を考慮し、かつこれらの２つの因子の優先度をユーザが指定できるため、ユーザの利用目的に合致したルールを生成できるという効果がある。

【０１９３】ルール生成処理では処理の進行状況が表示され、処理途中における中断処理も可能であり、かつ中断時点での途中結果が利用できる手段を設けているため、ルール生成処理に時間がかかるような場合に、ユーザの負担を減らし、作業効率を向上することが可能である。

【０１９４】また、記号値として与えられた項目の間の依存関係をデータ定義情報やデータ自身の分布から求め、ルール生成において冗長なルールを生成しないようにすることができるため、より単純でユーザの理解が容易なルールを、高速に生成することができるという効果がある。

【０１９５】さらに、生成したルールを用いて、記号値への変換方法に関するパラメータを学習により最適化できるため、精度がよく利用価値の高いルールを生成することができるという効果がある。

【図面の簡単な説明】

【図１】半導体不良分析における分析対象データの例を示す図

【図２】顧客分析における分析対象データの例を示す図

【図３】本実施例によるデータ分析方法の処理フロー図

【図４】データ分析システムの全体構成図

【図５】データテーブルの例を示す図

【図６】複数のデータテーブル間の関係定義の例を示す図

【図７】ルール生成処理仕様項目指定画面を示す図

【図８】項目自動選択処理フロー図

【図９】記号値変換方法一覧表示画面を示す図

【図１０】記号値変換方法指定画面を示す図

【図１１】メンバシップ関数の形状例を示す図

【図１２】ルール生成パラメータ指定画面を示す図

【図１３】ルール探索方法を説明する図

【図１４】ルール生成処理進行状況表示画面を示す図

【図１５】ルール生成処理フロー図

【図１６】学習によるメンバシップ関数形状変化の例を示す図

【符号の説明】

４０１…ユーザインタフェースモジュール、４０２…ルール生成処理実行モジュール、４０３…ルール学習処理実行モジュール、４０４…分析対象データベース、４０５…ルール生成用ワークファイル、４０６…ルール生成パラメータファイル、４０７…ルールファイル。

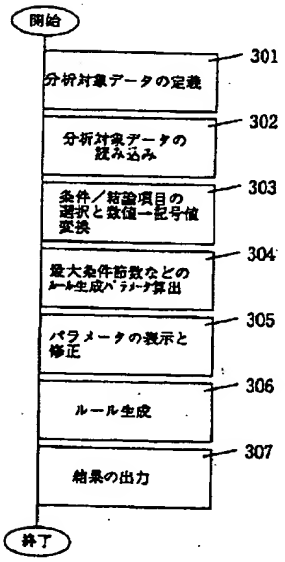
【図１】

マスクロット番号	洗浄方法	塗布方法	アルミ一層ショート	アルミ二層ショート	TH導通	…	電源不良	メモリ不良	総合不良発生率
TH5A0237	方法１	方法Ａ	5.1%	2.1%	0.2%	…	6.3%	32.5%	42.2%
TH5A0238	方法２	方法Ａ	0.2%	7.0%	0.0%	…	23.6%	12.2%	49.1%
TB1C0093	方法１	方法Ｂ	3.3%	3.5%	1.2%	…	4.1%	1.9%	6.8%
：	：	：	：	：	：	：	：	：	：

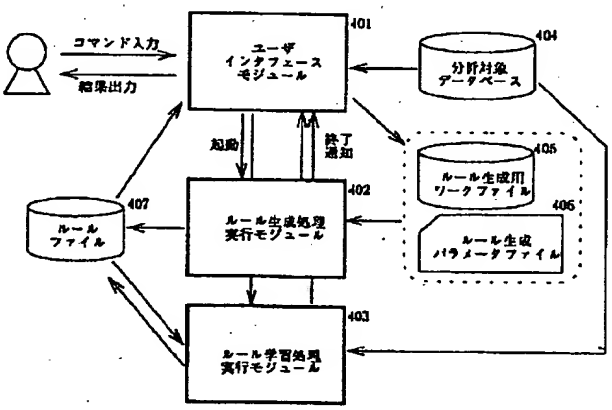
【図２】

顧客番号	氏名	年齢	性別	住所	支店コード	県コード	地域コード	預金残高	ローン残高	クレジットカード	…	商品購入
0158001	本多一郎	35	男	東京都世田谷区	108	25	3	250	1200	あり		あり
2048002	細川和夫	58	男	横浜市緑区	121	26	3	1200	510	あり		あり
3398024	朝倉知美	-	女	兵庫県芦屋市	257	37	5	120	340	なし		なし
1259842	上杉道彦	42	男	—	292	42	6	840	920	なし		あり
:	:	:	:	:	:	:	:	:	:	:		:
0912637	原田昌平	42	男	福岡県福岡市	359	46	7	780	0	あり		なし

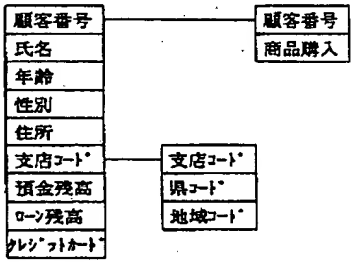
【図３】



【図４】



【図６】



【図５】

(a)顧客情報テーブル

顧客番号	氏名	年齢	性別	住所	支店コード	預金残高	ローン残高	クレジットカード
0158001	本多一郎	35	男	東京都世田谷区	108	250	1200	あり

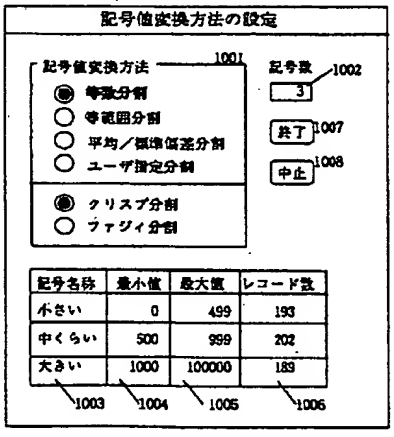
(b)支店情報テーブル

支店コード	県コード	地域コード	支店長氏名
108	25	3	松平明央

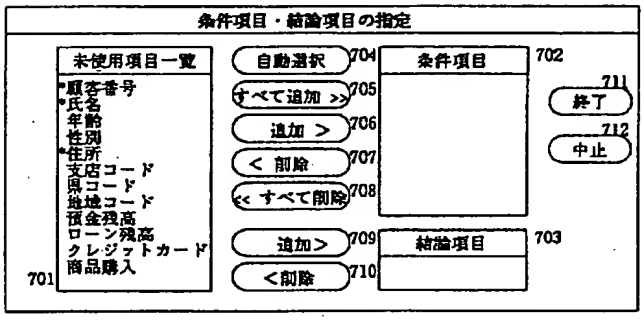
(c)商品購入履歴テーブル

顧客番号	商品購入
0158001	あり

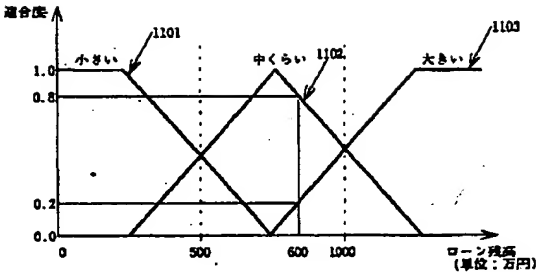
【図１０】



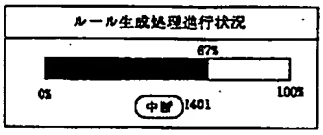
【図７】



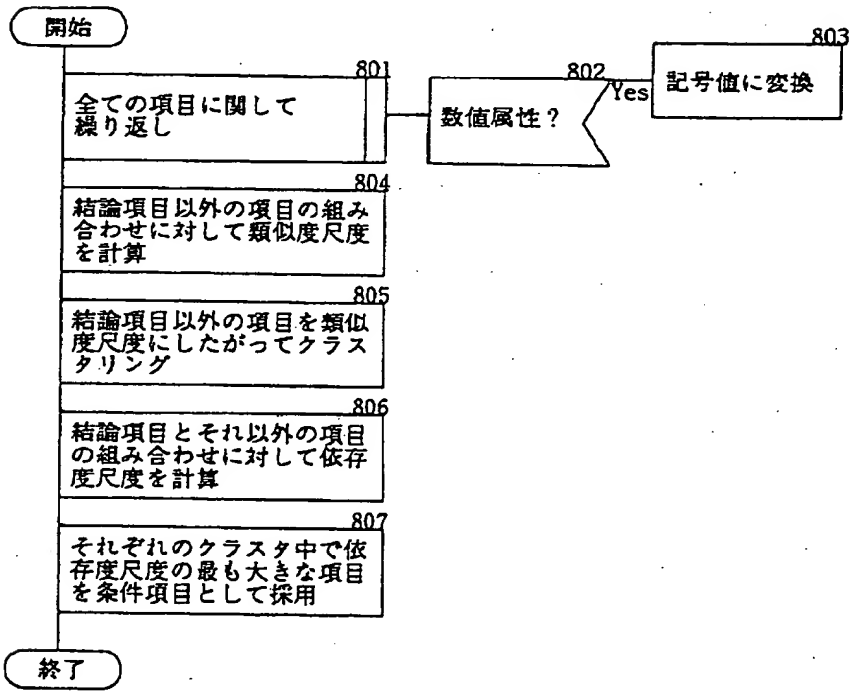
【図１１】



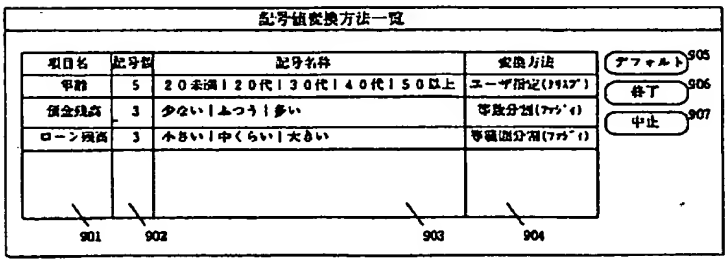
【図１４】



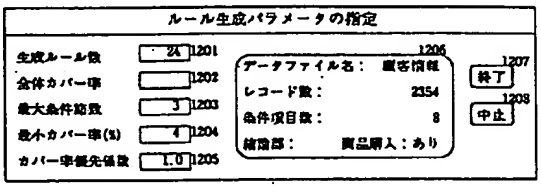
【図８】



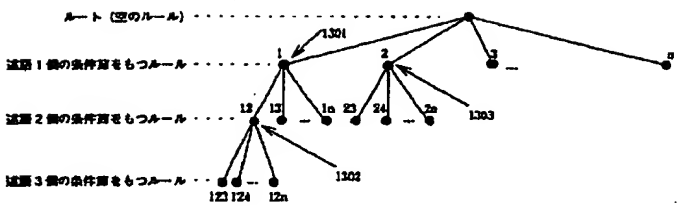
【図９】



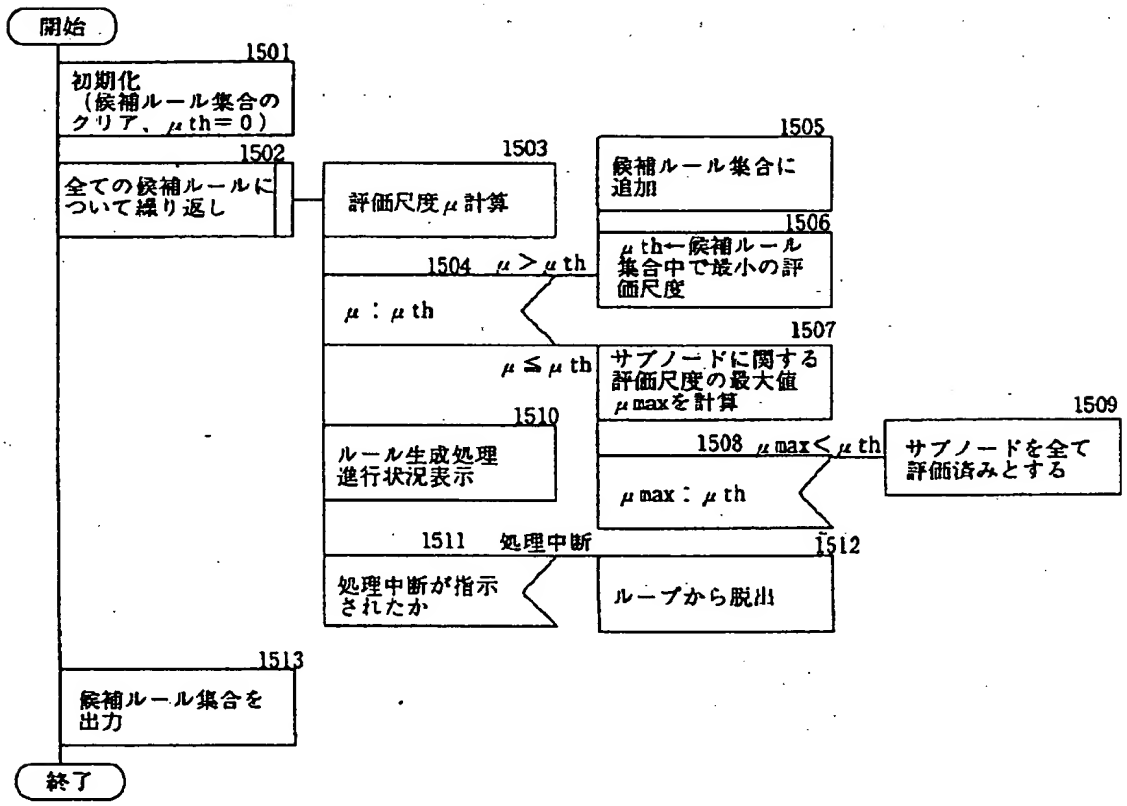
【図１２】



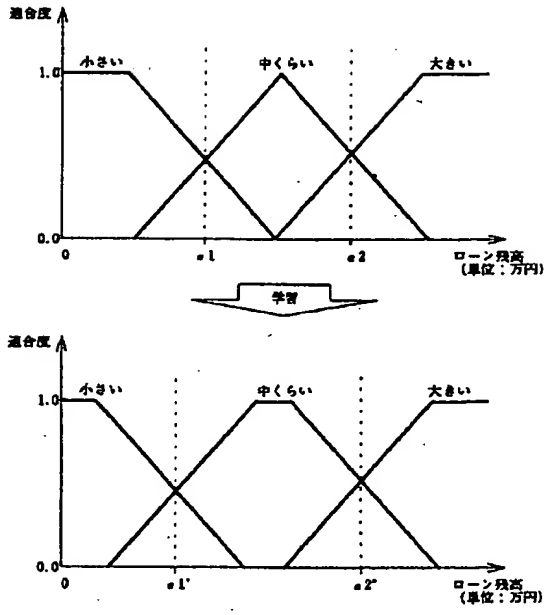
【図１３】


ルート（空のルール）
1301
1
2
3
n
1302
1303
123 124 12n

【図１５】


開始
1501
初期化
（候補ルール集合のクリア、μth＝０）
1502
全ての候補ルールについて繰り返し
1503
評価尺度μ計算
1504 μ＞μth
μ：μth
μ≦μth
1505
候補ルール集合に追加
1506
μth←候補ルール集合中で最小の評価尺度
1507
サブノードに関する評価尺度の最大値μmaxを計算
1508 μmax≦μth
μmax：μth
1509
サブノードを全て評価済みとする
1510
ルール生成処理進行状況表示
1511 処理中断
処理中断が指示されたか
1512
ループから脱出
1513
候補ルール集合を出力
終了

【図１６】



フロントページの続き

(72)発明者　伊藤　幸康
神奈川県横浜市戸塚区戸塚町5030番地　株式会社日立製作所ソフトウェア開発本部内

(72)発明者　髙橋　ヨリ
神奈川県横浜市戸塚区戸塚町5030番地　株式会社日立製作所ソフトウェア開発本部内